Large-scale Multi-modal Pre-trained Models: A Comprehensive Survey

Year: 2023 | Citations: 259 | Authors: Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiaoyong Wei

Abstract

With the urgent demand for generalized deep models, many pre-trained big models are proposed, such as bidirectional encoder representations (BERT), vision transformer (ViT), generative pre-trained transformers (GPT), etc. Inspired by the success of these models in single domains (like computer vision and natural language processing), the multi-modal pre-trained big models have also drawn more and more attention in recent years. In this work, we give a comprehensive survey of these models and hope this paper could provide new insights and helps fresh researchers to track the most cutting-edge works. Specifically, we firstly introduce the background of multi-modal pre-training by reviewing the conventional deep learning, pre-training works in natural language process, computer vision, and speech. Then, we introduce the task definition, key challenges, and advantages of multi-modal pre-training models (MM-PTMs), and discuss the MM-PTMs with a focus on data, objectives, network architectures, and knowledge enhanced pre-training. After that, we introduce the downstream tasks used for the validation of large-scale MM-PTMs, including generative, classification, and regression tasks. We also give visualization and analysis of the model parameters and results on representative downstream tasks. Finally, we point out possible research directions for this topic that may benefit future works. In addition, we maintain a continuously updated paper list for large-scale pre-trained multi-modal big models: https://github.com/wangxiao5791509/MultiModal_BigModels_Survey.